Def.: A "stochastic learning algorithm" is an assignment $S \mapsto \mu_S$, where $S \in (X \times Y)^n$ and μ_S is a prob. measure over a subset $\mathcal{F} \subseteq Y^X$ (with suitable σ-algebra, usually Borel).

We will write $R(\mu_S) := \mathbb{E}_{h \sim \mu_S}[R(h)]$, $\hat{R}_S(\mu_S) := \mathbb{E}_{h \sim \mu_S}[\hat{R}_S(h)]$

Exp.: Let r be a random variable with prob. dis $r \mapsto q(r)$ and $A_r : S \mapsto h_S$ a learning alg. for every r. Then we can define

$$\mu_S(H) := \int q(r)\, \mathbb{1}_{A_r(S) \in H}\, dr \quad \text{for any measurable } H \subseteq \mathcal{F}.$$

Then $R(\mu_S) = \mathbb{E}_r[R(h_S)]$, $h_S := A_r(S)$.

Def.: A stochastic learning alg. is "uniformly stable" with rate $\varepsilon : \mathbb{N} \to \mathbb{R}$, if $\forall n \in \mathbb{N}\ \forall S, S' \in (X \times Y)^n\ \forall (x, y) \in (X \times Y)$:

$$\mathrm{Ham}(S, S') = 1 \implies \left| \mathbb{E}\, L(y, h_S(x)) - \mathbb{E} L(y, h_{S'}(x)) \right| \leq \varepsilon(n)$$

where the expectation refers to the stochastic component of the learning alg. (e.g. $\mathbb{E} = \mathbb{E}_r$ in case $h_S = A_r(S)$).

Implications of stability carry over to stochastic alg.s. For instance, one easily obtains on-average generalization in the form

$$\mathbb{E}_S\left[\, R(\mu_S) - \hat{R}(\mu_S) \,\right] \leq \varepsilon(n)$$

The notion of differential privacy was introduced by C. Dwork (~2006) in the context of data-base analysis w.r.t. privacy. It turns out to imply stability:

Def.: A stochastic alg. $S \mapsto \mu_S$ is "ε-differentially private" if for all measurable $H \subseteq \bar{\mathcal{F}}$ and all S, S' with $\mathrm{Ham}(S, S') = 1$, we have:

$$\mu_S[h \in H] \leq e^{\varepsilon} \mu_{S'}[h \in H]$$

Cor.: If a learning alg. is ε-d.p., then it is $2(e^{\varepsilon} - 1)$ uniformly stable.

proof:

$$\left| \mathbb{E}\, L(y, h_S(x)) - \mathbb{E} L(y, h_{S'}(x)) \right| = \left| \int_h L(y, h(x)) \left[\mu_S(h) - \mu_{S'}(h) \right] dh \right|$$

Hölder ↓

$$\leq \underbrace{\sup_{h,x,y} |L(y, h(x))|}_{\leq 1} \int |\mu_S(h) - \mu_{S'}(h)|\, dh$$

$$\leq 2 \int_{h:\, \mu_S(h) \geq \mu_{S'}(h)} \mu_S(h) - \mu_{S'}(h)\, dh$$

$$\leq 2 \int \mu_{S'}(h) \left(\frac{\mu_S(h)}{\mu_{S'}(h)} - 1 \right) dh$$

$$\leq 2 \left(e^{\varepsilon} - 1 \right)$$

□